Abstract Title Page

Not included in page count.

Title: Synthesizing Evidence: Synthesis Methods for Evidence Clearinghouses

Authors and Affiliations:
Jeff Valentine
University of Louisville

Timothy Lau
University of Louisville

Abstract Body

Limit 4 pages single-spaced.

Background and Objectives

Following the theme of the first two presentations, this presentation will focus on the choices available for research synthesis when summarizing research evidence. For policymakers and practitioners who may be inclined to use the various evidence repositories as a resource for decision-making, how the evidence for a particular program or practice is summarized is important because different synthesis methods can result in different conclusions about program impacts. This presentation will review the synthesis practices available with a focus on the What Works Clearinghouse. The presenters will then discuss the strengths and limitations of different approaches to meta-analysis in a limited information environment, including unweighted meta-analysis, fixed effect and random effects meta-analysis using inverse variance weights, and Bayesian meta-analysis. After summarizing the various research synthesis options available, the presentation will use data from WWC intervention reports to show how the results differ across all of the synthesis options.

Research Design

The presenters will describe the current research synthesis practice of the What Works Clearinghouse (WWC) as well as several alternative models, including inverse-variance weighted fixed effect and random effects meta-analysis and Bayesian meta-analysis. The presenters will begin with an overview of the context of the WWC's intervention reports, with specific attention to the method used to synthesize the results of multiple studies and the context in which synthesis occurs. The data used for the illustration will be derived from published WWC intervention reports in which at least one synthesis across two or more studies was reported.

Data Collection and Analysis

When multiple studies that examine the same outcome are available, there is little debate that the best way to integrate the results of the studies is by using some form of meta-analysis. Below we discuss the WWC's approach to meta-analysis then outline the two most common forms of meta-analysis. We conclude with a discussion of Bayesian meta-analysis.

Unweighted meta-analysis. Current WWC synthesis practice is to conduct an unweighted meta-analysis. This is a fixed effect procedure that involves computing a simple average of the effects observed in the studies in the meta-analysis. While this procedure has its defenders (e.g., Shuster, 2010), it is uncommon and has significant drawbacks. Perhaps most important, there is little doubt that publication bias occurs. Publication bias refers to the tendency for studies lacking statistically significant finding on their primary outcomes to not be published. Studies that have been published are easier to find that unpublished studies – many of which never see the light of day. This means that small studies that by chance find large effects will tend to be overrepresented in a collection of studies (and conversely, small studies that observe effects that are not large will be underrepresented). In the presence of publication bias, the unweighted meta-analytic model will be biased against the null to a greater extent than any of the other common options. In other words, in most situations the unweighted model will produce effect sizes that are more inflated than the effect sizes that would be produced under another model. As such, the

WWC may be producing reports that give an unrealistically positive impression about the true effects of interventions.

A second limitation of the WWC's synthesis practice is that it does not involve reporting any information about the precision of the synthesized estimate. A typical WWC synthesis table will report each study's effect size, an interpretation of that effect size, and give an indication of the statistical significance of the effect size. Users are not, however, given an indication of the statistical significance of the synthesized effect, and as such are left to their own devices if they are interested in or could benefit from this information. In this situation, it would not be unusual for readers to adopt suboptimal rules. For example, they might assume that the synthesized effect is not statistically significant unless most or all of the component effects are statistically significant; this is a form of vote counting, the statistical properties of which are known to be poor (e.g., Hedges & Olkin, 1985).

Fixed effect meta-analysis. A more common meta-analytic model is known as fixed effect meta-analysis; this model is most often used in conjunction with inverse variance weights. The inverse variance weighting method essentially weights studies by their sample size (more technically, by a function of their sample size). This means that smaller studies get relatively less weight than larger studies. Among the popular approaches to meta-analysis, if one is willing to make the strong assumption that the studies are estimating the same population parameter, the fixed effect model represents the best <u>statistical</u> fit for the WWC and other similar efforts. This is because, if the strong assumption that studies are estimating the same population parameter is satisfied, the fixed effect model produces statistics that are more efficient than those produced by alternative methods.

Further, fixed effect meta-analysis is less vulnerable to publication bias than an unweighted meta-analysis. Because the inverse variance weighting method weights studies by their sample size, smaller studies get relatively less weight than larger studies; this is the mechanism by which the inverse variance weighting scheme tends to mitigate the effects of publication bias. In fact, for the meta-analyses that the WWC has conducted since 2010 that are based on exactly two studies, about 70% of the time the smaller study has observed the largest effect size (and overall, the correlation between sample size and effect size is about $r = -.15$). These related pieces of information are suggestive of a small degree of publication bias across these meta-analyses.

There is one conceptual limitation in implementing both the fixed effect model and the unweighted meta-analytic model in the context of the work of the WWC and other similar efforts. The fixed effect model is based on the assumption that the studies are estimating the same population parameter; any observed differences between studies are presumed to be a function of random sampling error; this statement is true of the unweighted approach as well. If this assumption is true, it suggests that if all studies in an unweighted meta-analysis or in a classical fixed effect meta-analysis were infinitely large, they would all obtain the same effect size. These illustrations imply that the unweighted and fixed effect models are most appropriate if the studies are highly similar to one another along important dimensions that contribute to variation in effect sizes (Hedges & Vevea, 1998) or in other words, if the studies are close replications of one another. Even though most clearinghouses impose constraints on the methods (via research standards) and on the characteristics of the intervention and sample (via protocols

developed in conjunction with content experts), most studies in the syntheses that public policy clearinghouses produce are probably not close replicates of one another. Instead, they likely are "ad hoc" replications that vary in known and unknown ways (Valentine et al., 2011). This means that neither the unweighted model nor the fixed effect model are well-suited to the kinds of syntheses these outfits produce.

Random effects meta-analysis. The most common meta-analytic model is known as the random effects model. Here, studies are thought of as estimating a distribution of studies that vary around a common effect size. The studies in a meta-analysis are therefore expected to vary from one another due to both known and unknown study characteristics in addition to random sampling error. In the usual inverse variance weighting scheme, studies are weighted by a function of their sample size and by an estimate of the extent to which the study estimates "disagree" with one another (the between-studies variance). Relative to the fixed effect model, the random effects model is generally more conservative. The confidence intervals arising from a random effects analysis will never be smaller and are usually larger than their fixed effects counterparts; this makes it less likely that the statistical conclusion following from an inferential test involving a random effects estimate will be a Type I error.

However, one critical limitation of this approach is that the estimate of one component of the analysis (the between studies variance, τ^2) is poor in most conditions under which the WWC will synthesize studies (i.e., when the number of studies is small). As a result, the estimated mean effect size and confidence intervals can be either too large or too small relative to what they "should" be, depending on whether the between studies variance is over- or under-estimated.

In summary, relative to more common alternatives the unweighted meta-analytic model almost certainly yields upwardly biased estimates of the mean effect size and implies a confidence interval that is almost always wider than would be produced under the common fixed effect analytic model. Both the unweighted model and the fixed effect model invoke a strong assumption – that the studies are highly similar to one another – and this assumption is probably not true in most syntheses produced by public policy clearinghouses. The random effects model provides a better conceptual fit that these models, but statistically, the estimate of the between studies variance is poor and as such statistically the random effects model is not a good choice for the clearinghouse-based syntheses as they are currently conducted.

Bayesian meta-analysis. One option is that clearinghouses such as the WWC could consider adopting a different statistical framework. Classical statistics is based on a theory that views parameters (unknown constants) as fixed. The implication is, for example, that while we would like to be able to say "There is a 95 percent chance the parameter is in this confidence interval," we can't because the parameter value is either in the interval or it is not. The "95 percent" is behavior in repeated sampling, not with respect to any one study. On the other hand, Bayesian statistics views parameters as random variables with distributions. These distributions tell how much information there is about the value of the parameter.

A second difference between classical and Bayesian statistics follows from the first: Each parameter has a prior distribution; that is, a distribution of the beliefs about the parameter before gathering data (in our case, before doing the meta-analysis). This prior distribution can be

informative, reflecting prior knowledge, or uninformative, reflecting prior ignorance. The information (if any) in the prior distribution is combined with the information in the data (expressed in the usual form of the likelihood, as in classical statistics) to produce a posterior distribution reflecting beliefs about the parameter(s) after the analysis.

As a result of these philosophical differences with classical statistics, the Bayesian framework provides advantages in user interpretation and in the use of prior information. The interpretations of important components of a study's results are more natural than those of classical statistics – this is an especially true and critical point for users who are not statisticians. Furthermore, the use of a somewhat informative prior may be justifiable in cases where we have both statistical and subject matter knowledge that suggests that the variability of true effect sizes across studies is not large. In the case of small meta-analyses, this makes inference much more informative than if we had no prior information.

Results

Using the 56 syntheses with at least two studies from existing What Works Clearinghouse intervention reports, the presenters will synthesize the data five ways: unweighted meta-analysis, classical fixed effect meta-analysis, classical random effects meta-analysis, Bayesian fixed effect meta-analysis, and Bayesian random effects meta-analysis. The following parameters will be compared:

1. A comparison of the means from each of the synthesis procedures, focusing specifically on how the unweighted mean performs relative to other options
2. A comparison of the confidence intervals from each of the syntheses (including the unweighted procedure's implied confidence intervals)
3. Empirical estimates of τ^2 across WWC reports (using several methods)
4. Descriptive statistics

Conclusions

Given that most of the syntheses that the public-policy clearinghouses do involve only a very small number of studies, they could consider not doing any syntheses at all unless the number of studies reaches some minimum threshold (e.g., five studies). However, consumers really would like to know something about the typical effect, even if the number of studies is small, and are likely to adopt idiosyncratic rules to help them do this. Alternatively, the clearinghouses could take a different approach to synthesis either in place of or as an adjunct to current practice. For example, the focus of WWC intervention reports, which tend to be about a relatively narrowly defined intervention, such as a brand name reading curriculum (like *Reading Recovery*) prevents the WWC from using a classical random effects model. An alternative would be to synthesize studies at a more abstract level, for example, by reviewing all supplemental reading interventions instead of reviewing each intervention type, like *Reading Recovery,* separately. If there are a sufficient number of studies to allow for good statistical power, moderator tests could be done to investigate the effects of specific components (e.g., total amount of supplemental instruction time, characteristics of individuals providing instruction, etc.). In addition, a Bayesian approach may also provide information on intervention effects that is affords easier interpretation by consumers.

Appendices

Not included in page count.

Appendix A. References

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis.* NY: Academic Press.

Hedges, L. V., & Vevea, J. L. (1998). Fixed-and random-effects models in meta-analysis. *Psychological Methods*, *3*, 486-504.

Shuster, J. J. (2010). Empirical vs natural weighting in random effects meta-analysis. *Statistics in Medicine, 29*, 1259-1265.

Valentine, J. C., Biglan, A., Boruch, R. F., Castro, F. G., Collins, L. M., Flay, B. R., Kellam, S., Mościcki, E. K., & Schinke, S. P. (2011). Replication in prevention science. *Prevention Science, 12*, 103-117.